Pokhara University

Faculty of Science and Technology

Course Code.: CMP 336	Full marks: 100
Course title: **Data Science and Machine Learning (3-1-2)**	Pass marks: 45
Nature of the course: Theory & Practical	Time per period: 1 hour
Level: Bachelor	Program: BE Software

1. Course Description

This course provides a comprehensive introduction to the fields of Data Science and Machine Learning, aimed at equipping students with the essential knowledge and practical skills required to analyze data, interpret data, apply machine learning methods and visualize results.

It will include the following information:

- Covers a wide range of topics, including data pre-processing, statistical analysis, machine learning algorithms, model evaluation, and the application of these techniques to solve real-world problems.
- Delivery approach includes hands-on labs, case studies and deep understanding of how to leverage data to make informed decisions.

2. General Objectives

The course is designed with the following general objectives:

- To provide students with a foundational understanding of Data Science and Machine Learning.
- To familiarize students with techniques for cleaning, transforming, and visualizing data to uncover patterns and insights.
- To provide the knowledge for use of mathematics such as statistics, probability for data analysis and machine learning,supervised learning algorithms, linear regression, decision trees, and support vector machines, and their applications.

- To expose students to unsupervised learning techniques such as clustering and dimensionality reduction, and their use in identifying patterns and simplifying data.

.

3. Contents in Detail

This section contains the details to be taught under the course.

Specific Objectives	Contents
● Intends to provide a brief introduction to the field of Data Science and Machine Learning. ● Learn about various domains within Data Science and how they interrelate. ● Helps students understand the significance of data in modern decision-making.	**Unit I: Introduction to Data Science and Machine Learning (4Hrs)** 1.1 Definition and Overview of Data Science and Machine Learning 1.2 Applications of Data Science in various industries 1.3 Types of Data: Clean Data and Dirty Data 1.3 Data Science, AI, and Machine Learning
● Intends to get students well-acquainted with data collection methods and preprocessing techniques. ● Able to apply various preprocessing techniques to clean and prepare data for analysis.	**Unit II: Data Collection and Preprocessing (7 Hrs)** 2.1 Different Data Collection Methods for Machine Learning: Surveys, Sensors, Web Scraping, APIs, Databases 2.2 Data Quality Issues: Missing Data, Noisy Data, Inconsistent Data, Data Transformation 2.3 Techniques for Handling Missing Data 2.4 Data Cleaning Techniques: Handling Outliers, Dealing with Categorical Data, Normalization, and Standardization 2.5 Dependent and independent variables
● Intends to provide students with the skills to explore and understand data. ● Learn about various EDA techniques to identify patterns and insights in the data	**Unit III: Exploratory Data Analysis (6Hrs)** 3.1 Introduction to EDA 3.2 Descriptive Statistics: Mean, Median, Mode, Standard Deviation, Variance, Skewness, Kurtosis 3.3 Data Visualization Techniques: Histograms, Box Plots, Scatter Plots, Heatmaps 3.4 Identifying Trends: Mann–Kendall, Spearman's Rank, Sen’s Slope 3.5 Correlations 3.6 Introduction to Hypothesis Testing
● Intends to provide students with the skills to explore and understand data. ● Learn about various EDA techniques to identify patterns and insights in the data	**Unit IV: Data Engineering (5 Hrs)** 4.1 Data pipeline, Design and Monitoring 4.2 Extract, Transform and Load (ETL) 4.3 Feature Engineering 4. 5 Feature Selection 4.6 Dimensionality Reduction: PCA, LDA
● Helps students learn how to implement basic machine learning models. ● Able to differentiate between various machine learning algorithms and their applications.	**Unit V: Introduction to Machine Learning (9 Hrs)** 5.1 Definition and Types of Machine Learning: Supervised, Unsupervised Learning, Reinforcement Learning 5.2 Overview of the Machine Learning

	5.3 Supervised Learning: Linear Regression, Logistic Regression, Decision Trees, Random Forest, k-NN, Support Vector Machines (SVM) 5.4 Unsupervised Learning: k-Means Clustering, Hierarchical Clustering methods Key Concepts: Training, Testing, Validation, Overfitting, Underfitting
● Apply basic and machine learning methods for detecting anomalies.	**Unit VI: Anomaly Detection (4 Hrs)** 6.1 Definition 6.2 Types: point, contextual, collective 6.3 Applications 6.4 Techniques for Anomaly Detection 6.4.1 Statistical Methods 6.4.2 Distance-based Methods 6.4.3 Density-based Methods 6.4.4 Clustering based Methods 6.4.5 Common Methods (one-class classification, isolation forest) 6.5 Anomaly Detection in High-Dimension
● Intends to get students well-acquainted with model evaluation techniques. ● Able to make use of various optimization techniques to improve model performance.	**Unit VII: Model Evaluation and Optimization (6 Hrs)** 7.1 Confusion Matrix, 7.2 Evaluation Metrics 7.2.1 Supervised: Accuracy, Precision, Recall, F1 Score, ROC Curve, AUC, MSE, True Positive Rate, False Positive, MSE, MAE, RMSE 7.2.2 Unsupervised: Purity, Rand Index, Silhouette Coefficient, Dunn Index 7.3 Cross-Validation Techniques 7.4 Hyperparameter Tuning: Grid Search, Random Search 7.5 Model Selection Techniques: Bias-Variance Trade-off, Ensemble Methods (Bagging and Boosting) 7.6 SMOTE Technique to Handle Imbalance 7.7 Time & Space Complexity of Machine Learning Models
● Helps students understand the ethical implications and legal considerations in data science.	**Unit VIII: Ethical and Legal Considerations in Data Science (4 Hrs)** 8.1 Data Privacy and Security 8.2 Ethical Issues in Data Science: Bias, Transparency, Accountability 8.3 Legal Considerations: Data Protection Laws, Intellectual Property

4. Methods of Instruction

The course will utilize a mix of lectures, tutorials, case studies, and lab sessions to support learning. Lectures will deliver core knowledge, while tutorials and case studies will enhance comprehension. Lab sessions will provide hands-on experience, enabling students to apply theory to practical, real-world situations. This integrated approach ensures a well-rounded learning experience, fostering both theoretical insight and practical skills essential for success in data science and analytics.

5. Case Studies

Students will complete the following case studies and submit their reports:

- Exploratory Data Analysis (Agricultural Commodities): Students will conduct a comprehensive exploratory data analysis on a dataset related to agricultural commodities. This will involve analyzing trends, patterns, and correlations to provide insights.
- Supervised Learning (Customer Churn Prediction in Telecommunications): Students will build and evaluate a supervised learning model to predict customer churn in the telecommunications industry. The case study will require them to preprocess data, select relevant features, and apply classification algorithms to identify customers at risk of leaving.
- Anomaly Detection in Real-World Applications: Students will implement anomaly detection techniques to identify unusual patterns or outliers in a real-world dataset. This case study will involve applying various anomaly detection methods to solve practical problems such as fraud detection or system monitoring.

Students are required to submit a detailed report documenting their approach, results, and analysis.

6. List of Tutorials

The following tutorial activities of 15 hours per group of maximum 24 students should be conducted to cover all the required contents of this course.

S.N.	Tutorials
1	● Using libraries of your programming choices (e.g. pandas, R) to manipulate datasets. ● Conducting exploratory data analysis (EDA) on real-world datasets. ● Cleaning and preprocessing data to prepare for modeling.
2	● Solving problems related to descriptive statistics (mean, median, mode, variance). ● Applying probability concepts to data science problems. ● Working with probability distributions and sampling techniques.
3	● Solving problems involving matrix operations and vector calculus.
4	● Applying linear algebra concepts to data transformations. ● Implementing supervised models like linear regression, decision trees, and k-nearest neighbors. ● Implementing unsupervised model like k-means, hierarchical

	● Implementing anomaly detection for real world data. ● Understanding the concept of overfitting and underfitting through practical examples. ● Hyperparameter tuning and model evaluation techniques.
6	● Creating visualizations using Matplotlib and Seaborn. ● Visualizing complex datasets and interpreting the results. ● Building dashboards using tools like Plotly or Dash.
7	● Implementing a complete machine learning pipeline from data collection to model deployment. ● Working on real-world datasets and competitions (e.g., Kaggle). ● Understanding the ethical implications and bias in machine learning.

7. Practical Works

S.N.	Practical works
1	Conduct an exploratory data analysis (EDA) on a public dataset.
2	Perform data manipulation tasks such as filtering, grouping, and summarizing.
3	Implement and compare different statistical techniques to analyze sample data (e.g., hypothesis testing, regression analysis).
4	Clean and preprocess a messy dataset (e.g., handling missing data, encoding categorical variables, feature scaling).
5	Implement different supervised learning algorithms (e.g., linear regression, decision trees) on a dataset.
6	Apply clustering techniques (e.g., K-means, hierarchical clustering) on a dataset and evaluate the clusters.
7	Perform a probabilistic model.
8	Apply anomaly detection methods in real world dataset.

8. Evaluation system and Students' Responsibilities

Evaluation System

In addition to the formal exam(s) conducted by the Office of the Controller of Examination of Pokhara University, the internal evaluation of a student may consist of class attendance, class participation, quizzes, assignments, presentations, written exams, etc. The tabular presentation of the evaluation system is as follows.

External Evaluation	Marks	Internal Evaluation	Marks
Semester-End Examination	50	Class attendance and participation	5
		Lab, Case study and Viva	15
		Internal Term Exam	30
Total External	50	Total Internal	50
Full Marks 50+50 = 100			

Students' Responsibilities:

Each student must secure at least 45% marks in the internal evaluation with 80% attendance in the class to appear in the Semester End Examination. Failing to obtain such a score will be given NOT QUALIFIED (NQ) and the student will not be eligible to appear in the End-Term examinations. Students are advised to attend all the classes and complete all the assignments within the specified time period. If a student does not attend the class(es), it is his/her sole responsibility to cover the topic(s) taught during the period. If a student fails to attend a formal exam, quiz, test, etc. there won't be any provision for a re-exam.

9. Prescribed Books and References

Text Book

Grus, J. *Data Science from Scratch: First Principles with Python*, Second Edition, O'Reilly Media.

Geron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow,* Second Edition, O'Reilly Media.

An Introduction to Statistical Learning by Gareth James et al.

O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishing Group.

Aggarwal, C. C. (2017). *Outlier analysis* (2nd ed.). Springer.

Reference Books

McKinney, W. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, Second Edition, O'Reilly Media.